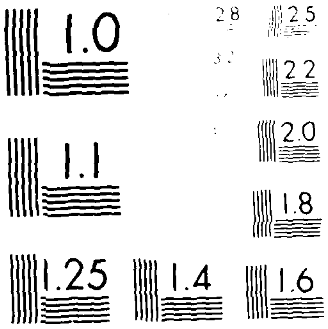
AD-A170 030   DETECTION OF OUTLIERS IN MULTIVARIATE LINEAR REGRESSION   1/1
MODEL(U) PITTSBURGH UNIV PA CENTER FOR MULTIVARIATE
ANALYSIS D N NAIK APR 86 TR-86-11 AFOSR-TR-86-0359
UNCLASSIFIED   F49620-85-C-0008   F/G 12/1   NL


1.0
2.8
2.5
3.2
2.2
2.0
1.1
1.8
1.25
1.4
1.6

Detection of Outliers in Multivariate Linear Regression Model

Dayanand N. Naik*
Center for Multivariate Analysis
University of Pittsburgh

Center for Multivariate Analysis

University of Pittsburgh

Detection of Outliers in Multivariate Linear Regression Model

Dayanand N. Naik*
Center for Multivariate Analysis
University of Pittsburgh

Technical Report No. 86-11

April 1986

Center for Multivariate Analysis
Fifth Floor, Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

*Part of the work of this author was sponsored by the Air Force Office of Scientific Research (AFOSR), under Contract F49620-85-C-0008 The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Detection of Outliers in Multivariate Linear Regression Model

SUMMARY

In this article we suggest multivariate kurtosis measure as a statistic for detection of outliers in a multivariate linear regression model. The statistics has some local optimal properties.

Some key words: Multivariate linear regression model, Detection of outliers, Multivariate kurtosis, Locally best invariant test.

1. INTRODUCTION

Several authors have dealt with the problem of detection of outliers in linear model. See Cook and Weisberg (1982). However, the corresponding multivariate problem is difficult and there is not much work in that area. For excellent entensive surveys of the outlier literature see Barnett and Lewis (1984). In this paper we give a locally optimum procedure for detection of outliers based on Mardia's (1970) multivariate sample kurtosis. Result is based on extension of Ferguson's (1961) work to multivariate case on the similar lines of Sinha (1984) and Schwager and Margolin (1982). The idea of using Ferguson's (1961) work on outlier detection, with suitable modifications to linear regression problems, was suggested by C.R. Rao. The multivariate problem is an offshot of that idea.

2. NOTATIONS AND REDUCTION OF THE PROBLEM

Consider the multivariate linear regression model

$$\underset{n\times p}{Y} = X\underset{m\times p}{B} + E \quad , \quad \text{rank}\ (X) = m \qquad (2.1)$$

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
[illegible] SR-TR- 86-0359

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
DETECTION OF OUTLIERS IN MULTIVARIATE LINEAR REGRESSION MODEL

5. TYPE OF REPORT & PERIOD COVERED
Technical - April 1986

6. PERFORMING ORG. REPORT NUMBER
86-11

7. AUTHOR(s)
Dayanand N. Naik

8. CONTRACT OR GRANT NUMBER(s)
F49620-85-C-0008

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Center for Multivariate Analysis
515 Thackeray Hall
University of Pittsburgh, Pittsburgh, PA 15260

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research
Department of the Air Force
Bolling Air Force Base, DC 20332

12. REPORT DATE
April 1986

13. NUMBER OF PAGES
8

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Multivariate linear regression model, detection of outliers, multivariate kurtosis, locally best invariant test.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
In this article we suggest multivariate kurtosis measure as a statistic for detection of outliers in a multivariate linear regression model. The statistics has some local optimal properties.

REFERENCES

BARNETT, V. & LEWIS, T. (1984). Outliers in Statistical Data, 2nd Edition, J. Wiley & Sons, N.Y.

COOK, R.D. & WEISBERG, S. (1982). Residuals in Influence in Regression, Chapman & Hall.

FERGUSON, T.S. (1961). On the Rejection of Outliers. Proceedings of IV Berkeley Symposium, 1, 253-287.

MARDIA, K.V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. Biometrika, 57, 519-530.

SCHWAGER, S.J. & MARGOLIN, B.H. (1982). Detection of Multivariate Normal Outliers, Ann. Statist., 10, 943-954.

SINHA, B.K. (1984). Detection of Multivariate Outliers in Elliptically Symmetric Distributions, Ann. Statist., 12, 1558-1565.

THEIL, H. (1965). The Analysis of Disturbances in Regression Analysis, J. Amer. Statist. Assn, 60, 1067-1079.

VAIDYA, H. (1985). Some Contributions to Familial Data Analysis, Ph.D. thesis, University of Pittsburgh, USA.

WIJSMAN, R.A. (1967). Cross-sections of Orbits and Their Application to Densities of Maximal Inariants. V Berkeley Symp., 1, 389-400.

Assume rows of E to be independent, each distributed as $N(0,\Sigma)$, i.e. $\mathrm{Vec}(E) \sim N(0,\Sigma \otimes I_n)$. We write (2.1) in the form

$$(Y_1:\ldots:Y_p) = (X\beta_1:\ldots:X\beta_p) + (\varepsilon_1:\ldots:\varepsilon_p) \tag{2.2}$$

The BLUE of β is $\hat{\beta}_i = (X^1X)^{-1}X^1Y_i$, $i=1,2,\ldots p$. The residual vectors are

$$\hat{\varepsilon}_i = Y_i - X\hat{\beta}_i\ ,\ i=1,2,\ldots,p$$

Thus we have $\hat{E} = (\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_p)$ and $\mathrm{Vec}(\hat{E}) \sim N(0,\Sigma \otimes Q)$ where $Q = I - X(X^1X)^{-1}X^1$. An unbiased estimate of Σ is $S = \hat{E}^1\hat{E}/(n-m)$.

Let us denote n row vectors of p × 1 dimension by $e_1, e_2,\ldots,e_n$. If one or more of the quadratic forms

$$e_iS^{-1}e_i,\quad i=1,2,\ldots,n$$

are unusually large, then we identify corresponding observations as outliers. In the following we adopt the procedure due to Theil (1965) to get uncorrelated residual vectors, keeping the problem at hand in mind. First, we order the quadratic forms $e_iS^{-1}e_i$, $i=1,2,\ldots,n$ in the increasing order of magnitude. Then, rewrite the model (2.1) starting with the row having smallest $e_i'S^{-1}e_i$ and continuing until the observation vector with largest $e_i'S^{-1}e_i$ is at the bottom.

For notational convenience let us take the rewritten model to be the same as (2.1). Now Theil's (1965) BLUS method involves choosing X_0 from X, starting with the first row, so that X_0^{-1} exists.

Then (2.1) can be written as

$$\begin{pmatrix} Y_m \\ Y_{(n-m)} \end{pmatrix} = \begin{pmatrix} X_0 \\ X_1 \end{pmatrix} B + E \qquad\qquad \text{or}$$

$$\begin{pmatrix} Y_m : \ldots : Y_{mp} \\ \\ Y_{(n-m)1} : \ldots : Y_{(n-m)p} \end{pmatrix} = \begin{pmatrix} X_0\beta_1 : \ldots : X_0\beta_p \\ \\ X_1\beta_1 : \ldots : X_1\beta_p \end{pmatrix} + E \tag{2.3}$$

Now make the transformation,

$$U_i = Q_{11}^{\frac{1}{2}} Y_{(n-m)i} - Q_{11}^{\frac{1}{2}} X_1 X_0^{-1} Y_{mi} \text{ , } i=1,2,\ldots,p \tag{2.4}$$

where $Q_{11} = I - X_1(X'X)^{-1}X_1'$ and $Q_{11}^{\frac{1}{2}}$ is such that $Q_{11} = Q_{11}^{\frac{1}{2}} Q_{11}^{\frac{1}{2}}$. Each U_i , $i=1,2,\ldots,p$ is $(n-m) \times 1$ residual vector and has the property that, if $\mathcal{U} = (U_1,\ldots,U_p)$ then

$$\text{vec}(\mathcal{U}) = N(0, \Sigma \otimes I_{n-m}).$$

That is to say, rows of $\mathcal{U}$ are independently distributed as $N(0,\Sigma)$, the p-variate normal distribution. Thus, we have (n-m) i.i.d. observations from a p-variate normal distribution with mean zero and covariance matrix Σ, and we want to detect whether there are any outliers among them. Similar problem for observations from $N(\mu,\Sigma)$ for unknown μ has been solved by Schwager and Margolin (1982) and Sinha (1984).

3. FORMULATION OF THE PROBLEM AND MAIN RESULT

Let X be $n \times p$ observation matrix, such that rows of X are independent and each row is a p-variate normal with mean 0 and covariance matrix Σ. Possibility of outliers with mean slippage can be incorporated by considering the model

$$X = \Delta A \Sigma^{\frac{1}{2}} + Z\Sigma^{\frac{1}{2}} \tag{3.1}$$

with Δ a nonzero scalar, $A = (a_{ij})$ an arbitrary $n \times p$ matrix such that some of the rows of A are zero and Z, mean zero, unit variance, independent normal variables. Unless $\Delta = 0$, the observation X_i corresponding to the ith row of X is an outlier if the ith row of A is nonzero.

The general outlier problem then consists of the model (3.1) and the null hypothesis $H_0:\Delta = 0$ versus the alternative $H_1:\Delta \neq 0$. We derive locally optimum test of H_0 Vs H_1 employing invariance arguments through the use of a group of transformation keeping the testing problem invariant.

The above testing problem is invariant under the action of the group $G = P \times Gl(p)$ where P denotes the group of all $n \times n$ permutation matrices with element Γ_α, $Gl(p)$ the group of $p \times p$ nonsingular matrices with elements C. The group operations are defined by (1) post multiplication of X by any nonsingular matrix $C \varepsilon Gl(p)$ and (2) permutation of the rows of X by premultiplying X by $\Gamma_\alpha \varepsilon P$. Without loss of generality assume $\Sigma = I$.

The following lemma due to Wijsman (1967) is taken from Sinha (1984).

Lemma 3.1 Let $h(x/\Delta)$ be the pdf of x, let $T = t(x)$ be a maximal invariant under the transformation G and let P_Δ^T be the distribution induced by T under Δ. Then the pdf of T w.r.t. P_0^T evaluated at $T = t(x)$ is given by

$$\frac{dp_\Delta^T}{dp^T} = \frac{\int_G h(g \cdot x/\Delta)|C'C|^{n/2} d\nu(g)}{\int_G h(g \cdot x/\Delta=0)|C'C|^{n/2} d\nu(g)} \tag{3.2}$$

where ν is left invariant measure on G. Here $g \cdot x = \Gamma_\alpha x C$, $\Gamma_\alpha \varepsilon P$, $C \varepsilon Gl(p)$ and $\nu = \nu_1 \times \nu_2$, ν_1 is discrete uniform probability measure with mass $1/n!$ at each of the $n!$ elements $\Gamma_\alpha \varepsilon P$ and

$$d\nu_2(C) = dC/|C'C|^{p/2}.$$

Lemma 3.2 The ratio in (3.2) reduces to

$$\frac{\Sigma_\alpha \int_{Gl(p)} \operatorname{etr} - \frac{1}{2}\{C'C - 2\Delta C'S^{-\frac{1}{2}}(\Gamma_\alpha X)'A + \Delta^2 A'A\}|C'C|^{\frac{n-p}{2}} dC}{\Sigma_\alpha \int_{Gl(p)} \operatorname{etr}(-\frac{1}{2} C'C)|C'C|^{\frac{n-p}{2}} dC} \qquad (3.3)$$

Proof is easy proceeding on the similar lines as in Sinha (1984).

Now we proceed to evaluate the expression in (3.3). An exact evaluation of the expression is not necessary to evaluate locally best invariant test. We use Taylor series expansion upto a few terms evaluated at $\Delta = 0$. Making a transformation from C to -C, it is clear from (3.3) that the ratio of the pdf's depend only on Δ^2. Let N_Δ and N_0 be the numerator and the denominator of (3.3) respectively. We assume the conditions for taking derivative inside the integral signs hold. Then, using Taylor expansion we write

$$N_\Delta = N_0 + N_0^1\Delta + N_0^{(2)}\frac{\Delta^2}{2!} + N_0^{(3)}\frac{\Delta^3}{3!} + N_0^{(4)}\frac{\Delta^4}{4!} + \ldots.$$

$$= N_0 + N_0^{(2)}\frac{\Delta^2}{2!} + N_0^{(4)}\frac{\Delta^4}{4!} + \ldots.$$

Using the results (Lemma 4.1) of Schwager and Margolin (1982) we can easily show that coefficient of $\frac{\Delta^2}{2!}$,

$$-\operatorname{tr}(A'A)N_0 + \Sigma_\alpha \int_{Gl(p)} [t\, C'S^{-\frac{1}{2}}(\Gamma_\alpha X)'A]^2\, e^{-\frac{1}{2}\operatorname{trc}'c}|C'C|^{\frac{n-p}{2}} dC,$$

is a constant. The coefficient of $\frac{\Delta^4}{4!}$ apart from a constant is

$$\Sigma_\alpha \int_{Gl(p)} [\operatorname{tr} AC'S^{-\frac{1}{2}}(\Gamma_\alpha X)']^4\, e^{-\frac{1}{2}\operatorname{tr}C'C}|C'C|^{\frac{n-p}{2}} dC \qquad (3.4)$$

Let $T(x) = b_{2,p} = n \sum_{i=1}^{n} (X'S^{-1}X_i)^2$ be multivariate kurtosis measure defined as in Mardia (1970). Let L(A) be such that

$$n(n-1)\ L(A) = (n-2) \sum_{i=1}^{n} ||r_i||^4 - 3(\sum_{i=1}^{n} ||r_i||^2)^2 \tag{3.5}$$

where $||r_i||^2 = a_i' a_i$, a_i is the i^{th} row of A.

Now (3.4) apart from a constant can be written, using the results due to Ferguson (1961), Schwager and Margolin (1982) and Sinha (1984), as

$$c_1 T(x)\ L(A) + C_2. \tag{3.6}$$

Then we have the following theorem.

Theorem For the outlier problem discussed, the locally best invariant test of $H_0{:}\Delta = 0$ Vs $H_1{:}\Delta \neq 0$ conditional on A, is: if $L(A) > 0$, reject H_0 whenever $b_{2,p} \geq k$; if $L(A) < 0$, reject H_0 whenever $b_{2,p} \leq k'$. The constants k,k' are determined by the size of the test and L(A) is the function of A given in (3.5).

Proof Application of Lemma 3.2 and the generalized Neyman-Pearson Lemma along with (3.6) completes the proof of the theorem.

One can use asymptotic distribution of $b_{2,p}$, obtained by Mardia (1970), to find the cutoff points k,k'. Or else, in specific problems, one can use simulation to compute k,k'.

Now returning back to the multivariate regression model considered in section 2; we test the hypothesis $\Delta = 0$ Vs $\Delta \neq 0$ using the uncorrelated residual vectors obtained in (2.4) and applying the above theorem. If the hypothesis is rejected then we identify the observation corresponding to the largest $e_i'S^{-1}e_i$, as an outlier. Removing the outlier observation frcm the data, further testing can be done for more outliers.

We would like to remark that the kurtosis measure is very sensitive for the presence of outliers and hence is a very useful tool for detection of outliers. This fact, at least in the case of univariate regression models, was realized in a data analysis problem considered by Vaidya (1985).

ACKNOWLEDGEMENTS

This work is based on a part of author's Ph.D. dissertation written under the direction of Professor C.R. Rao at University of Pittsburgh. It is a pleasure to thank professor Rao for his many valuable suggestions and comments.

END
DTIC
8-86